View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Aaltodoc Publication Archive

Aalto University
School of Business

Negative words in 10-K's and filing period stock returns

Bachelor's Thesis
Elmeri Niemelä
Aalto University School of Business
Department of Finance

Fall 2019

Author Elmeri Niemelä

Title of thesis Negative words in 10-K's and filing period stock returns

Degree Bachelor's degree

Degree programme Finance

Thesis advisor(s) Theresa Spickers

Year of approval 2019 **Number of pages** 15 **Language** English

Abstract

Previous research has developed a word list specifically designed to classify the negative tone of financial texts. Research has shown evidence that classifying 10-K documents by frequency of negative words can predict returns after the filing date. Using this negative word list, I analyze if the return predictability patterns still emerge. In order to control for implementation differences, I collect two samples of 10-K documents with daily stock data: from 1994 to 2008 as the control sample and a new sample from 2008 to 2018. Negative tone analysis with the frequency based method is not sufficient enough, to consistently predict stock returns. Results from previous research using this method can only be replicated in their original time frame from 1994 to 2008.

Keywords textual analysis, 10-K, bag-of-words, sentiment analysis, filing period returns, word lists

Table of Contents

1. Introduction

1.1. Contribution to existing literature

With the exponential growth of available computing power during the last century, new methods of textual analysis have become available. From a financial perspective, the question of interest is that can these methods be used to predict stock value. In this paper, I will try to reproduce promising results from previous research with a newer set of data. My goal is to see if previously documented effects are still relevant a few years after the initial study has been published.

In a pioneering paper called 'When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks' (2011) Loughran and McDonald discuss the effects of using different word lists to measure the tone of text. They suggest that the common word list used in negative tone classification called 'Harvard-IV-4 TagNeg file' misclassifies most of the words used in financial context such as *costs*, *taxes* or *liabilities*. These words are negative in common texts whereas in a financial context they are considered a neutral representation of the status of a business. Additionally, they find proof that misclassifications in the Harvard word list could introduce type one errors to negative tone analysis. If the word 'cancer' is classified as negative, industry segments such as health sector are biased towards more negative tone.

To correct these misclassifications Loughran and McDonald build their own 'FinNeg' word list of 2,337 words that they deem to have negative implications in a financial sense. Using this word list they classify negativity of 10-K documents from 1994 to 2008 and match them to available financial data. They show that there is a strong pattern between negative word frequencies and negative stock returns on a 4-day period after the 10-K document has been filed. In this study, I reproduce the same pattern in the same time period and then continue to test if the pattern appears consistently with a new data sample from 2008 to 2018. Thus, when referencing the "original study" in the text, I am referring to this study made by Loughran and McDonald in 2011.

I follow the steps of Loughran and McDonald and construct a multivariate regression model to test the relation of negative tone in 10Ks to company stock returns. First, I run the model against the same time period as Loughran and McDonald, from 1994 to 2008. As expected, the results show a strong pattern between the high frequency of negative words in 10-K filing and decreased stock returns after the filing date. However, after performing the same procedure on the new data

sample from 2008 to 2018, the effect of decreased returns disappears. As the original study was published in 2011, it is possible that professional investors have implemented a trading strategy that mitigates the effect. But as noted by the original authors, the model has insignificant practical value due low explanatory power. In the concluding chapter, I consider the flaws of textual analysis methodology as a more the likely explanation.

1.2. Research on textual analysis

Textual analysis has a long history starting from the 13th century where it was used to index common phrases in biblical translations. In the survey of Textual analysis in accounting and finance (2016), Loughran and McDonald provide an excellent overview of how textual analysis has grown especially during the last decades. The amount of methods available for researchers is growing rapidly. New machine learning methods such as “Deep learning” and “Cloud Robotics” (Pratt 2015) deepen the possibilities of textual analysis. As access to the internet provides a bigger than ever corpus for textual analysis, the subject becomes increasingly relevant in most disciplines. In accounting and finance, a particular point of interest is the Securities and Exchange Commission (SEC) filings, which can easily be downloaded and are free to research. As all public US companies are required to make regular SEC filings, this provides a large corpus for financial text analysis.

Textual analysis in finance is considered to be a form of qualitative analysis that is commonly divided into analyzing readability, targeted phrases, topic modeling, measures of document similarity and sentiment analysis which is the focus of this study (Loughran and McDonald, 2016). The words used by company executives have been shown to reflect the future performance of the company stock returns. According to Loughran and McDonald, groundbreaking papers by Frazier, Ingram, and Tennyson (1984), Antweiler and Frank (2004), Das and Chen (2007), Tetlock (2007), and Li (2008) have inspired finance researchers to study the impact of this qualitative information on stock returns. “Can we tease out sentiment from mandated company disclosures and contextualize quantitative data in ways that might predict future valuation components?”, this important question asked by Loughran and McDonald (2016) is potentially answered by textual analysis.

Sentiment analysis in finance is most often focused on analyzing the impact of negative and positive word lists. Positive word lists have little success so far, most likely due to common negation of positive words. Businesses often frame bad news using positive word negation such as “profits did not grow” but are less likely to do so when conveying good news such as “losses did

not increase" (Loughran and McDonald 2011). For this reason, I only focus on examining the impact of negative words.

1.3. Bag-of-words method

The most common method of textual analysis used in finance is called bag-of-words and it's the method I use in this study as well. The goal is to break the text down and quantify it so it can be used in statistical analysis. The first step is to tokenize the text into separate words. Then I use the FinNeg word list developed by Loughran and McDonald (2011) to identify words that are classified as negative. In this step all the stop words such as 'the', 'he' and 'do' are removed since they do not contain any meaningful information about the text. Usually, the next step is including all the inflections of the root word by either stemming or lemmatizing the tokenized words. In this study, stemming or lemmatizing is not needed since the word list already explicitly contains inflections that do not change the meaning of the word. Once I have categorized the words into 'bags' based on negativity, I simply count the frequency related to the total amount of meaningful words in the document.

One distinctive problem with the bag-of-words methodology is immediately imminent: The context of the words is ignored and they are treated as independent units. This is a major flaw in the method since words in natural languages are highly dependent on the context they are presented in. Given the tool set available today, calculating word counts seems like just the beginning (Loughran and McDonald 2016).

The implementation details from the bag-of-words process are crucial in textual analysis. Distilling meaningful information about text into numbers is inherently imprecise and vague descriptions of the parsing process make replication of the study close to impossible. For this reason, as suggested by Loughran and McDonald (2016), links to all the details and the exact implementation done in this study are found in the appendix 'Variable Definitions and Internet Resources'.

Although bag-of-words is a common word categorization method used in finance, some researches have used different approaches based on vector distance, Naïve Bayes classifications, likelihood ratios, and other classification algorithms. Loughran and McDonald (2011) reference Das and Chen (2001), Antweiler and Frank (2004), and Li (2009) as example papers using these alternative methods.

2. Data and methods

Table I: 10-K Sample Creation

	Old Sample 1994-2008 (N = 40032)		New Sample 2008-2018 (N = 25460)	
Source/Filter	Sample Size	Observations Removed	Sample Size	Observations Removed
Full 10-K Document EDGAR 10-K/10-K405 2008-2018 complete sample	121994		91086	
Include only first filing in a given year	121275	719	90011	1075
CRSP PERMNO match	81886	39389	58671	31340
CRSP market capitalization data available	65677	16209	44464	14207
Price on filing date day minus one is greater than $3	57570	8107	38410	6054
Returns for day 0-3 event period	57523	47	38374	36
At least 60 days of volume prior to file date (used to caluclate turnover)	54167	3356	36403	1971
Book-to-market COMPUSTAT data available and book value greater than zero	40384	13783	25554	10849
Atleast 2,000 words in 10-K	40033	351	25468	86
Rows containing any missing values	40032	1	25460	8

2.1. Sample creation compared to Loughran and McDonald's original study

I follow the steps of Loughran and McDonald (2011) by downloading all of the 10-Ks from the EDGAR website (www.sec.gov). In order to control for implementation differences compared to the original study, I collect two samples: the old sample, from 1994 to 2008, same as used by Loughran and McDonald, and the new sample from 2008 to 2018.

Table I shows how the old and the new samples are filtered to meet my financial data requirements for the regression analysis. I focus on discussing the older sample since I can compare it to the original study, but the same points apply to the newer sample as well. The biggest data filter is the CRSP PERMNO match that removed 39389 observations from the older sample.[1] This is expected since according to Loughran and McDonald most of the firms with missing PERMNO's are real estate, non operating, or asset-backed partnerships/trusts that are required to file with the SEC.

Price and market capitalization data are required to be available exactly one day before the filing date. This requirement filters more observations than in the original study with 16209

1 I use Wharton Data Services 'CRSP/ COMPUSTAT Merged Database - Linking Table' to link CIK numbers to CRSP PERMNOs

observations removed compared to 5834.[2] Stocks that have a price less than 3 dollars are removed to reduce the role of bid-ask bounce. (Loughran & McDonald, 2011). To calculate share turnover I require at least 60 observations of volume to be available in the period [-252, -6] prior to the filing date. Another notably large filter is the availability of book-to-market COMPUSTAT data and book value being greater than zero. This requirement removes 13783 observations compared to 4770 in the original study.

For the older sample from 1994 to 2008, these criteria yield a sample size of 40032 observations compared to 50115 in the original study and 7133 unique firms compared to 8341. Although the older sample is smaller than in the original study, I can still reproduce the original findings correctly. I follow the same procedure for the newer sample from 2008 to 2018 yielding a sample size of 25460 observations and 4320 unique firms.

2.2. Parsing the 10-K documents

In order to turn a body of text into analyzable numbers, some parsing is required. Parsing is done in two phases. A substantial amount of the documents consists of technical data such as HTML-code, images and other ASCII-encoded data that is not in text form. Phase one consists of cleaning out those non-English parts from the documents. Exhibitions and tables are also removed as they are less likely to contain tonal indicators, and instead have larger proportions of noisy template language (Loughran and McDonald 2011). I follow the procedure explained in detail on McDonald's web page and based on that I create my own parsing implementation.[3]

In the second phase, I parse the 10-K document into frequencies of words in order to construct a commonly used text classification model called bag-of-words[4]. I use the model to classify documents by the percentage of negative words that they contain. I do this phase with exactly the same implementation as in the original study by Loughran and McDonald (2011) since the code is available at the University of Notre Dame's website.

---

2 The difference is most likely due to not filtering the data firstly based on the stock being reported as an ordinary common stock equity firm, as Loughran and McDonald do in their study.

3 McDonald's description of the document cleaning procedure is from here: https://sraf.nd.edu/data/stage-one-10-x-parse-data/#_ftn1. My own parsing implementation is found here: https://github.com/thecodebasesite/bachelor/blob/master/clean_sec_files.py

4 This common method is introduced in the introduction section 1.3. Common pitfalls of this method are discussed in conclusions.

2.3. Variables

To evaluate the economic impact of negative word frequencies, the dependent variable for all of the regressions is excess stock returns. Excess stock return is defined as a 4 day holding period return on the stock (starting from the filing date) minus the CRSP value-weighted buy-and-hold market index return over the 4-day event window. Buy and hold returns are calculated as a geometric average of daily returns during the 4 day period. The event period of 4 days, chosen by Loughran and McDonald (2011), is based on paper by Griffin (2003, Table II).

Control variables include firm size, book-to-market, share turnover and a dummy variable for NASDAQ listing. I leave out Fama–French alpha and institutional ownership from my model, which causes a slightly lower R^2 than in the original study, but regardless I reproduce the original findings correctly. In addition, I include industry dummies and a constant in each regression. I use the same 48-industry classification scheme of Fama and French (1997), as used Loughran and McDonald, to control for cross-sectional effects in the data. I use SIC codes from the CRSP data to convert into Fama and French industries.[5] Robustness of the model is tested by checking how much the results vary by reducing and changing the selection of the control variables.

5 Implementation of my SIC to Fama and French Industry mapping is found here: https://github.com/thecodebasesite/bachelor/blob/master/sic_mapping.py

3. Results

Table III: Summary Statistics for the samples

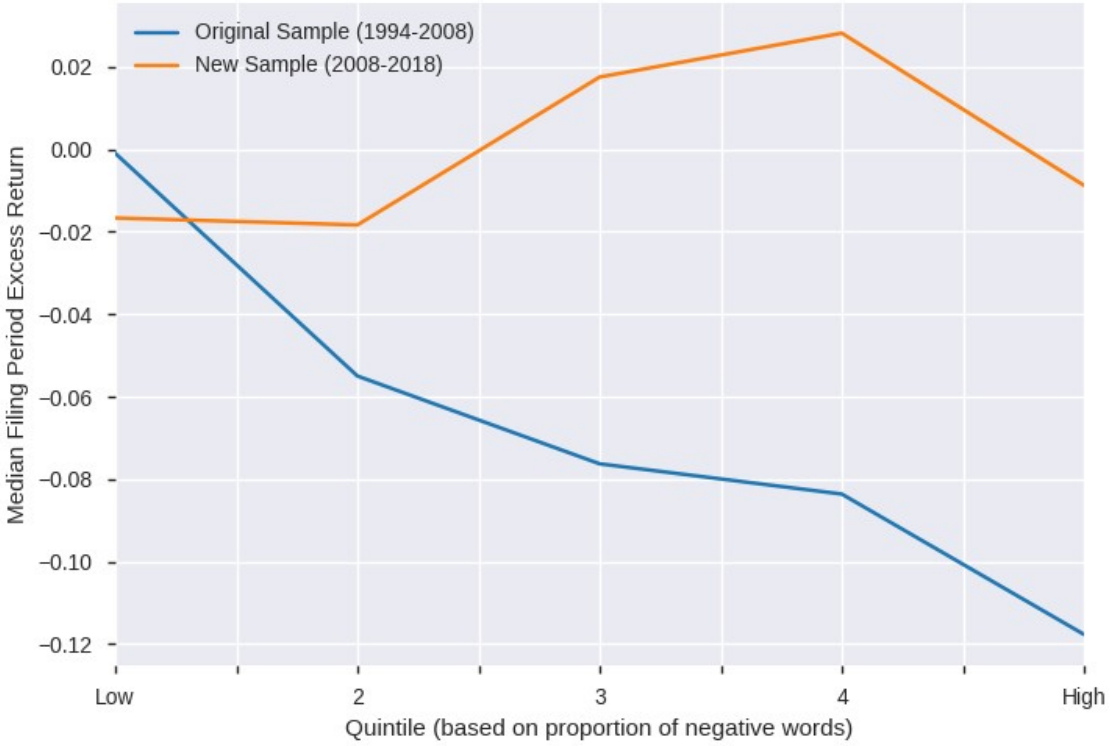
	Old Sample 1994-2008 (N = 40032)			New Sample 2008-2018 (N = 25460)		
Variable Name	**Mean**	**Median**	**Standard Deviation**	**Mean**	**Median**	**Standard Deviation**
Negative Word Frequency	1.30%	1.26%	0.54%	1.78%	1.76%	0.43%
Event period [0, 3] excess return	-0.12%	-0.07%	1.83%	-0.01%	-0.00%	1.77%
Size (billions $)	2.94$	0.34$	15.17$	6.20$	0.92$	25.55$
Turnover	1.63	1.04	2.42	2.36	1.74	4.15
Book-to-market	0.56	0.47	0.41	0.55	0.45	0.41
NASDAQ Dummy	52.93%	100.00%	49.91%	51.87%	100.00%	49.97%

3.1. Sample Description

Summary statistics for both samples are reported in Table III. I examine a total of 3.4 billion words across both samples. Significant variable differences between the old and new samples are firm size and turnover, both of which have increased in the newer sample. Increase in average firm size is not a surprising development, especially since the prices are not inflation corrected. Increased share turnover can be explained by increases in institutional trading and more widespread use of quantitative trading strategies (Chordia, Roll, Subrahmanyam, 2008).

The most interesting change is in event period excess returns that have converged to the mean of -0.01% in the new sample compared to -0.12% in the old sample. The new sample contains less negative returns overall compared to the old sample, which relates to the different results of these samples. A hint on growing amount of word misclassifications is that even though the average negative word frequency has risen from 1.30% to 1.78%, the average event period excess returns have increased instead of decreasing.

Figure 1. Median filing period return by quintile for both samples.



3.2. Filing period returns and negative word frequencies

For both of the samples, I do the same examination of the market's reaction at the time of 10-K filing, as Loughran and McDonald (2011). If the bag-of-words model is still relevant, I expect decreased returns around the filing date with 10Ks that have a high frequency of negative words. Figure 1 reports the median filing period excess returns by quintiles of negative word frequencies. Each quantile contains the same amount of observations, with the first quintile having the lowest frequency of negative words and the last having the highest. Lines are drawn for both the original sample from 1994 to 2008 and for the newer sample from 2008 to 2018. The original sample shows clearly the pattern that was documented by Loughran and McDonald in their study (Figure 1, 2011). As expected, firms with more negative words in 10-Ks experience more negative returns after filing the 10-K document. However, with the new sample, the pattern disappears completely. There is no consistent relation between filing period excess returns and 10-K negative word frequencies after the year 2008 where Loughran and McDonald's sample ended.

Next, I examine the relation of negative tone and excess returns in a multivariate context using multiple control variables. The dependent variable in each regression is the event period excess returns expressed as a percent. I follow the same Fama-MacBeth (1973) methodology as Loughran and McDonald (2011) by first grouping the data into quarterly cross-sections and then running

regressions on all the companies in each quarter. The estimates for each regression are saved and weighted by frequency, since most of the filing dates are clustered around the first two quarters (Griffin 2003). To test if the results are significantly different from zero, I do a t-test that controls for Newey-West (1987) standard errors with one lag. Tables IV and V report the regression results for the old and the new sample respectively.

Table IV: Negative word frequencies using filing period excess returns regressions. (1994 - 2008 sample)

	Coefficient	Standard Error	T-Value
Variable name			
Negative Word Frequency	-0.001325	0.000474	-2.794386
Log(size)	0.000775	0.000240	3.221438
Log(turnover)	-0.001065	0.000522	-2.039621
Log(book-to-market)	0.001956	0.000502	3.893302
NASDAQ Dummy	-0.000327	0.000459	-0.711623

Average Adjusted R-Squared: 1.27 %

Table V: Negative word frequencies using filing period excess returns regressions. (2008 - 2018 sample)

	Coefficient	Standard Error	T-Value
Variable name			
Negative Word Frequency	0.000399	0.000649	0.615295
Log(size)	0.001024	0.000326	3.143577
Log(turnover)	-0.001017	0.000506	-2.008974
Log(book-to-market)	0.001865	0.000593	3.142581
NASDAQ Dummy	0.001222	0.000940	1.299777

Average Adjusted R-Squared: 2.73 %

For both of the samples, the regression coefficients of negative word frequency tell the same story as Figure 1. Assuming a higher negative word frequency has a negative impact on stock returns, the regression coefficient should be negative. This is not the case with the newer sample, having a positive coefficient and statistically insignificant t-statistics of 0.615. Meanwhile the older sample produces a negative coefficient that is statistically significant. The t-statistic of -2.79 in the older sample is not far from -2.64 produced be Loughran and McDonald (2011, Table IV). For the older sample I get a slightly lower R^2 of 1.27% compared to 2.52% in the original study since I did not include Fama–French alpha and institutional ownership. Nevertheless, higher proportions of negative words in 10-Ks are associated with lower excess returns, but only up to 2008.

As noted by Loughran and McDonald, even with the old sample, the model has diminishing practical value as the R^2 of 2.52% in the original study is relatively low. And even that a small amount of explanatory power is lost when the new sample is examined. “Textual analysis is not the ultimate key to the returns cipher.” as stated by Loughran and McDonald in their original study (2011).

4. Conclusions

I started with the results from Loughran and McDonald (2011) in their paper called ‘When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks’. They show evidence that negative tone inside 10-Ks is related to negative excess returns on a 4-day period after the document is filed. I show that the effect is reproducible, but only in their specific time frame from 1994 to 2008. After 2008 the effect has completely disappeared due to combination of multiple reasons that are discussed in the next chapter.

4.1. Possible pitfalls

As the original study was published in 2011, professional traders may have already implemented a trading strategy based on the study that mitigates the effect. But with an R^2 of only 2.52%, there is not that much to gain by implementing a trading algorithm based on the bag-of-words approach. More probable explanation is in the primitiveness of the methodology.

A problematic example related to SEC filings HTML formatting is given by Loughran and McDonald (2016). Potomac Electric Power uses of <TABLE> tags to define text paragraphs instead of numeric tables in its 20040312 filing. Since numbers do not convey tone, the contents of every <TABLE> tag are removed in the first phrasing phase meaning that this document is incorrectly parsed. Additionally, the amount of attachments such as PDF files has grown substantially with the newer filings which opens the possibility that the crucial information related to tone is no longer available in plain text format.

As discussed in section 1.3, the bag-of-words methodology used in the study has major flaws, most importantly the assumption of word independence from context. Given the amount of template language and other noise incorporated by 10-K filings, it is very likely that more sophisticated models are needed for sentiment analysis of 10-Ks. Fast development of machine learning methodology opens new possibilities for future research in this area.

According to Loughran and McDonald, a property associated with any normal distribution called Zipf's law (section 1.4.3 of Manning and Schütze, 2003), potentially causes certain misclassified words to add a large bias on the results. The same effect is sometimes refereed as the 80-20 rule: 80% of negative tone is produced by 20% of the most common negative words. Figure 1 from Loughran and McDonald (2016), shows clearly how the word counts in SEC filings are dominated by few common words. Misclassification of these common words can change the results substantially.

4.2. Future research

The common methodologies of financial text analysis used in this study are most likely just scratching the surface. Machine learning methods such as "Deep learning" and "Cloud Robotics" (Pratt 2015) may be used to incorporate the context and to capture the deeper meaning of financial texts. Future research on term weighting can provide a more structured and objective way to reduce the possibility of major misclassifications (Loughran and McDonald, 2016). Developing standards around textual analysis research in finance will reduce the bias of methodology decisions done by researchers.

Appendix: Variable Definitions and Internet Resources

Variable Definitions

These definitions follow Loughran and McDonald's study (2011) as close as possible. The prefile date Fama–French alpha and institutional ownership are left out due to time constrains. With the exception of book-to-market, the definitions are identical to theirs. See Loughran and McDonald (2011, p. 63) 'Appendix: Variable Definitions'. Links to implementation details are below.

Size	The number of shares outstanding times the price of the stock as reported by CRSP on the day before the filing date
Book-to-market	Book value per share from COMPUSTAT Fundamentals Annual (WRDS library 'compd', file 'funda') where fiscal year equals filing date year. Divided by price of the stock on the day before the filing date. After removing negative values I winsorize the variable at 1% level.
Share turnover	The volume of shares traded in days [−252,−6] prior to the file date divided by shares outstanding on the file date. At least 60 observations of daily volume must be available to be included in the sample.
NASDAQ dummy	A dummy variable set equal to one for firms whose shares are listed on the NASDAQ stock exchange, else zero.

Internet Resources

1. Homepage for all of the code used in the study:

   https://github.com/thecodebasesite/bachelor

2. Data profiling links in the README.md file found on the bottom of the homepage.

3. Variable implementations:

   https://github.com/thecodebasesite/bachelor/blob/master/data_organize.py

References

Antweiler, W. and Frank M., 2004, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards., Journal of Finance, Vol. 59, No. 3, pp. 1259–1294

Chordia, T., Roll R., and Subrahmanyam, A., 2011, Recent trends in trading activity, Journal of Financial Economics, Vol. 101, pp.243–263.

Das, S. R. and Chen, M. Y., 2007, Yahoo! for Amazon: Sentiment Extraction from Small Talk on theWeb. Management Science Vol. 53, No. 9, pp. 1375–1388.

Fama, E.F. and French, K.R., 1997, (cited in Loughran, T. and McDonald, B.. 2011 p. 42) Industry costs of equity, Journal of Financial Economics 43, Pages 153–193.

Fama, E.F. and MacBeth J., 1973, (cited in Loughran, T. and McDonald, B.. 2011 p. 52) Risk, return, and equilibrium: Empirical tests, Journal of Political Economy 81, 607–636.

Frazier, K. B., Ingram, R. W. and Tennyson, B. M., 1984, A Methodology for the Analysis of Narrative Accounting Disclosures. Journal of Accounting Research, Vol. 22, No. 1, pp. 318–331.

Griffin, P., 2003, (cited in Loughran, T. and McDonald, B.. 2011 p. 41) Got information? Investor response to Form 10-K and Form 10-Q EDGAR filings,Review of Accounting Studies 8, 433–460.

Li, F, 2008, Annual Report Readability, Current Earnings, and Earnings Persistence. Journal of Accounting and Economics Vol. 45 pp. 221–247.

Loughran, T. and McDonald, B., 2011, 'When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks', The Journal of Finance, Vol. 66, No. 1, pp. 35-66

Loughran, T. and McDonald, B., 2016, Textual Analysis in Accounting and Finance: A Survey, Journal of Accounting Research, Vol 54, Issue 4, pp. 1187-1230

Manning, C. D. and Schütze H., 2003, Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press

McDonald, B., 2019, Stage One 10-X Parse Data, University of Notre Dame, viewed 14 November 2019, <https://sraf.nd.edu/data/stage-one-10-x-parse-data/#_ftn1>

McDonald, B., 2019, Code, University of Notre Dame, viewed 14 November 2019, <https://sraf.nd.edu/textual-analysis/code/>

Newey, W.K. and West K.D., 1987, (cited in Loughran, T. and McDonald, B.. 2011 p. 52) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, Econometrica, Vol. 55, 703–708.

Pratt, G.A., 2015, Is a Cambrian Explosion Coming for Robotics?, Journal of Economic Perspectives, Vol. 29, No. 3, pp. 51–60

Tetlock, P. C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market., Journal of Finance, Vol. 62, 1139–1168.